SelfAdapt: Unsupervised Domain Adaptation of Cell Segmentation Models

Fabian H. Reith[1,2,4,5*], Jannik Franzen[1,3,4,5], Dinesh R. Palli[1,6],
J. Lorenz Rumberger[2,4,5†], Dagmar Kainmueller[3,4,5†]

[1]Charité - Universitätsmedizin, Berlin, Germany [2]Humboldt-Universität zu Berlin, Berlin, Germany
[3]Universität Potsdam, Digital Engineering Faculty, Potsdam, Germany
[4]Helmholtz Imaging
[5]Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany
[6]Ludwig-Maximilians-University, Munich, Germany

Abstract

Deep neural networks have become the go-to method for biomedical instance segmentation. Generalist models like Cellpose demonstrate state-of-the-art performance across diverse cellular data, though their effectiveness often degrades on domains that differ from their training data. While supervised fine-tuning can address this limitation, it requires annotated data that may not be readily available. We propose SelfAdapt, a method that enables the adaptation of pre-trained cell segmentation models without the need for labels. Our approach builds upon student-teacher augmentation consistency training, introducing L2-SP regularization and label-free stopping criteria. We evaluate our method on the LiveCell and TissueNet datasets, demonstrating relative improvements in $AP_{0.5}$ of up to 29.64% over baseline Cellpose. Additionally, we show that our unsupervised adaptation can further improve models that were previously fine-tuned with supervision. We release SelfAdapt as an easy-to-use extension of the Cellpose framework. The code for our method is publicly available at https://github.com/Kainmueller-Lab/self_adapt.

1. Introduction

Deep neural networks have become a widely applied approach for cell instance segmentation in microscopy data. In particular, generalist models such as Cellpose [20], StarDist [17], Mesmer [5] and Micro-SAM [1] have emerged as powerful segmentation tools, capable of handling a broad range of imaging modalities. These models are trained on diverse datasets and achieve state-of-the-art generalization performance. Nevertheless, their performance still often drops significantly on domains which differ from their training data [14, 18]. This limitation is particularly pronounced in biomedical imaging, where new data is often "out-of-domain" due to the wealth of cell types and tissues considered in biomedical research, highly diverse protocols for sample preparation, and continuously evolving imaging modalities – with potentially significant impact on model performance.

To address performance degradation caused by domain shifts, domain adaptation (DA) methodology seeks to adapt a model trained on a source domain to an unlabeled or sparsely labeled target domain. Traditional approaches for domain adaptation rely on supervised fine-tuning on the target domain [14], autoencoder pre- or auxiliary training [4, 6, 15] on source and target domain, and domain adversarial learning [3, 6]. However, these approaches require either labeled data from the target domain, or labeled data from the source domain, or at least image data from the source domain. This makes them unsuitable for adapting off-the-shelf models: First, acquiring labeled target annotations in biomedical applications is often prohibitively expensive and time-consuming, requiring expert knowledge and extensive resources; Second, off-the-shelf models usually do not ship with their training data.

This challenge is addressed by source-free unsupervised domain adaptation (UDA), which enables adaptation without any access to source data nor target labels. In particular, student-teacher frameworks with exponential moving average (EMA) updates [21] have shown strong potential for improving model generalization. These methods maintain a teacher model that is iteratively updated via EMA and used to generate pseudo-labels for a student model, allowing the network to learn from unlabeled samples. Techniques like

[*] Corresponding author: fabian.reith@mdc-berlin.de

[†] Equal contribution; when citing, it is permitted to change author order; author order was determined at random.

FixMatch [19] and ACTIS [16] further enhance this process by enforcing consistency between a student's predictions under different augmentations, extracting reliable training signals from unlabeled data. While originally showcased in semi-supervised scenarios where the UDA objective is complemented by small sets of labeled target data, Archit and Pape [1] show that these approaches can be adapted for pure source-free UDA scenarios in bio-medical applications; However, to do so, their method relies on a specialized architecture for uncertainty quantification [7] and thus does not lend itself to off-the-shelf pre-trained models.

To fill this gap, we propose SelfAdapt, a source-free UDA framework whose core training procedure is applicable to any model trained for pixel-wise classification- and/or regression tasks. We showcase this framework on cell instance segmentation, for which we also introduce and evaluate two distinct label-free stopping criteria: a general, representation-based metric (embedding distance) and an instance-specific metric (False Negative rate). SelfAdapt expands over ACTIS [16], which only considered models trained for pixel-wise classification tasks. By incorporating regression targets, SelfAdapt becomes applicable to many popular cell segmentation models, including Cellpose [14, 20], Stardist [17], and Mesmer [5]. We further expand over ACTIS by introducing L2-SP regularization [10] into the objective: L2-SP regularization constrains the adapted weights to stay close to their pre-trained values, thus serving to replace the regularizing effect of ACTIS's supervised component without compromising training stability.

We showcase SelfAdapt by applying it to Cellpose, arguably the most widely used off-the-shelf cell instance segmentation model, improving over the generalist baseline by large margins. Interestingly, our results even suggest improvements over off-the-shelf *fine-tuned* Cellpose models, i.e., models adapted to the target domains in a traditional supervised fashion [14].

In summary, our key contributions are:

1. SelfAdapt, an easy-to-use, label-free UDA method featuring a broadly applicable L2-SP-regularized training procedure and two distinct early stopping criteria: a general embedding-based metric and another tailored for instance-based tasks.
2. Seamless integration of SelfAdapt into the Cellpose framework for immediate deployment to the community.
3. Comprehensive evaluation on multiple cellular imaging datasets, demonstrating substantial performance improvements over generalist as well as fine-tuned Cellpose baselines.

2. Method

2.1. Augmentation Consistency Training

SelfAdapt builds upon ACTIS [16], a student-teacher framework [21] that enforces augmentation consistency, i.e., consistency between predictions on strongly and weakly augmented views of the input data, and employs teacher-generated pseudo labels [9, 19]. ACTIS further advances this training framework for cell instance segmentation by introducing a confidence-based loss masking scheme that filters uncertain predictions, combined with temporal ensembling [8] that improves pseudo-label quality through prediction averaging.

Following ACTIS, SelfAdapt's teacher model generates pseudo-labels by averaging predictions across multiple weakly augmented versions of an input image. Specifically, we apply flips and 90-degree rotations as weak augmentations, compute predictions for each transformation, and apply inverse transformations before averaging the outputs.

We further follow ACTIS in their confidence-based loss masking scheme, which applies to pixel-wise classification outputs. Expanding upon ACTIS, we also tackle pixel-wise regression outputs, using standard deviation across the different teacher predictions, i.e., a standard measure for regression uncertainty, as filtering criterion. For each scalar regression target, we maintain a moving average of the 80th uncertainty percentile to determine reliable regions. The student model, receiving strongly augmented inputs with intensity adjustments, contrast changes, blurring and Gaussian noise, is trained using a weighted combination of mean squared error for regression targets and cross-entropy loss for classification targets:

$$\begin{aligned}\mathcal{L}_{\text{self}} = &\mathcal{L}_{\text{CE}}(p_{\text{student}}, p_{\text{teacher}}) \cdot M_p \\ &+ \lambda_{bal} \cdot \mathcal{L}_{\text{MSE}}(f_{\text{student}}, f_{\text{teacher}}) \cdot M_f \,,\end{aligned} \tag{1}$$

where p represents class probability maps, f represents regression output predictions, M_p and M_f are binary masks excluding pixels with uncertainty above the threshold, and λ_{bal} balances the contribution of regression- vs. classification outputs. Only pixels with uncertainty below our dynamic threshold contribute to the loss, thus steering training to focus on reliable regions.

2.2. L2-SP Regularization

To preserve the features learned during pre-training while allowing for domain-specific adaptations, we employ L2-SP regularization [10] alongside standard weight decay. The L2-SP regularization term is defined as:

$$\mathcal{L}_{\text{L2-SP}} = \lambda_{L2\text{-}SP} \sum_i (w_i - w_i^0)^2 \,, \tag{2}$$

where w_i represents the current weights, w_i^0 denotes the initial pre-trained weights, and $\lambda_{L2\text{-}SP}$ controls the regulariza-

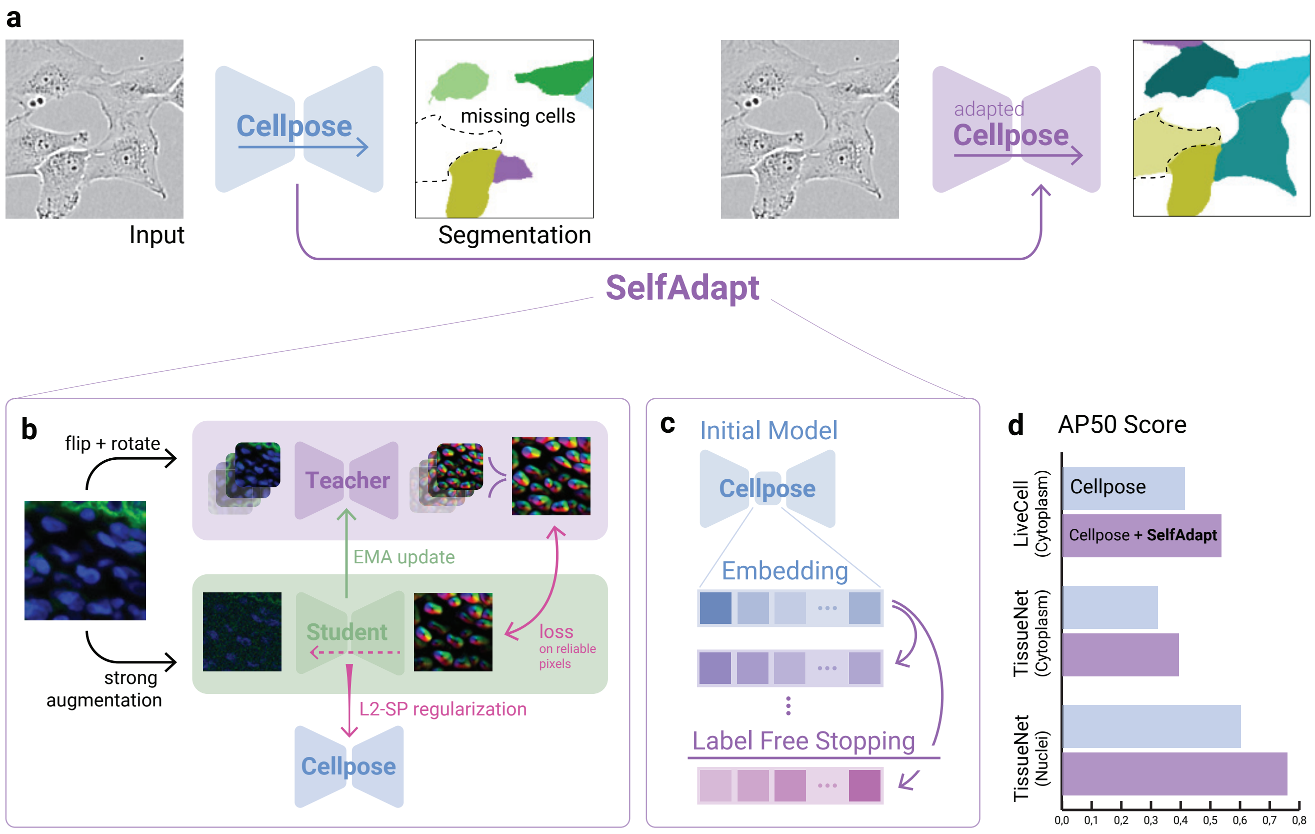

Figure 1. Overview of SelfAdapt. a) Generalist Cellpose model showing limited performance on a target domain, while our adapted model improves segmentation accuracy. b) Student-teacher framework with strong augmentations for the student and weak augmentations (flip+rotate) for the teacher, coupled with EMA updates and L2-SP regularization. c) Early stopping mechanism based on embedding distances from initial model. d) Performance comparison showing $\text{AP}_{0.5}$ improvements across datasets.

tion strength. This term explicitly encourages the model to maintain proximity to its initialization, serving to stabilize training in place of the supervised sub-objective of ACTIS.

2.3. Label-free Early Stopping Criteria

In fully unsupervised learning scenarios, determining the optimal stopping point is a critical challenge. While manual inspection is an option, it is subjective and not scalable. To address this, we propose an approach that relies on two label-free metrics to signal when to stop training.

The first is the False Negative (FN) rate, adapted from [13], which tracks the fraction of instances detected by the initial model (θ_0) that are missed by the current model (θ_t). It is defined as:

$$\text{FN}_{\text{rate}}(t) = \frac{\text{FN}(t)}{\text{TP}(t) + \text{FN}(t)}. \tag{3}$$

Here, $\text{TP}(t)$ are instances detected by both the initial and current models, while $\text{FN}(t)$ are instances detected initially but missed by the model at iteration t.

The second is the mean Euclidean distance (D_{emb}) between the current and initial bottleneck feature embeddings (E) across the validation set:

$$D_{emb}(t) = \frac{1}{N}\sum_{i=1}^{N} ||E(x_i;\theta_t) - E(x_i;\theta_0)||_2. \tag{4}$$

In this formulation, $E(x_i;\theta_t)$ is the bottleneck feature embedding for an input image x_i from the model with its current weights, which is compared against $E(x_i;\theta_0)$, the corresponding embedding from the initial, pre-trained model. This distance is then averaged over all N images in the validation set.

We selected these two metrics after a systematic investigation of a broader set of candidates, including other output-based metrics (e.g., prediction confidence, uncertainty) and representation-based metrics (e.g., proximity to source embeddings). A detailed analysis comparing these candidates and justifying our selection is presented in the Results section (Sec. 3.2).

Finally, the two selected criteria were calibrated on a representative dataset to establish general-purpose thresholds, which are used in all subsequent experiments.

3. Results

3.1. Experimental Setup

We evaluate SelfAdapt by adapting state-of-the-art Cellpose models on three challenging benchmarks: cytoplasm segmentation on the LiveCell dataset [2], and both cytoplasm and nuclei segmentation on TissueNet [5]. To ensure fair comparison, we train on the official training set and report performance on the test set for each benchmark. All experiments are run three times with different random seeds, and we report the mean and standard deviation of the Average Precision at an IoU of 0.5 (AP$_{0.5}$) [12]. We selected this metric as it is a common standard for cell instance segmentation and the primary metric used in the original Cellpose papers [14, 20], ensuring a direct and fair comparison with the baseline models.

Our implementation details are as follows: We train for 25,000 iterations with a batch size of eight. The AdamW [11] optimizer is used with a weight decay of 0.001. The learning rate follows a linear warm-up and a cosine annealing schedule. For our loss function, we set the balancing weight $\lambda_{bal} = 0.5$. The L2-SP regularization strength, $\lambda_{L2\text{-}SP}$, was tuned based on the specific model-dataset combination. For the results in Table 1, the values were set as follows: a strong regularization of $\lambda_{L2\text{-}SP} = 10^{-2}$ was used for the TissueNet (Nuclei) model and the fine-tuned TissueNet (Cytoplasm) model. A moderate value of 10^{-4} was used for the base models on LiveCell and TissueNet (Cytoplasm). Finally, a weaker regularization of 10^{-5} was applied when adapting the fine-tuned LiveCell model. The teacher model is updated via EMA with $\alpha = 0.99$. For cytoplasm and nuclei segmentation, we use the generalist 'cyto2' and 'nuclei' Cellpose models, respectively. For experiments on already fine-tuned models, we use the domain-specific checkpoints provided by Cellpose 2.0 [14].

3.2. Analysis and Selection of Early Stopping Criteria

As outlined in Sec. 2.3, our method relies on automated, label-free criteria to determine when to stop adaptation. To justify our choice, we performed a detailed analysis on a representative adaptation run (TissueNet Nuclei), comparing numerous candidate metrics against the true test set performance (the "oracle"). Our findings are summarized in Figure 2.

We first investigated metrics derived directly from the model's output. A naive approach is to monitor prediction confidence, such as the mean cell probability. However, this proved highly unreliable (Fig. 2a). We observed that confidence increased monotonically throughout training, continuing to rise even as the true performance degraded. We also explored more sophisticated uncertainty measures, such as the variance in predictions across multiple test-time augmentations (TTA). While these metrics were not always monotonic, their peaks and troughs did not correlate with the oracle performance peak, often providing a misleading signal for when to stop.

Next, we evaluated metrics based on the model's internal feature representations. One hypothesis was that the adapted target embeddings should move closer to the general feature space of the source domain. As shown in Fig. 2b, this proximity (measured by Euclidean distance) does indeed provide a smooth monotonic signal. However, this approach has two critical drawbacks: it violates our core source-free constraint by requiring access to the original training data, and as we will show, a source-free alternative provides an equally reliable signal.

This analysis led us to select two metrics that are both source-free and reliably track model drift from its stable, pre-trained state. The first is the False Negative (FN) rate (Fig. 2c), a self-consistency metric that measures the fraction of initially detected instances that the model "forgets" over time. The second is the mean Euclidean distance, D_{emb} (Fig. 2d), which measures the drift of the model's feature embeddings from their robust initial state. Both of these metrics provide predictable, monotonic curves that are ideal for setting a threshold, without needing access to labels or source data. We note that while the FN rate is effective, it is most reliable when the initial model does not suffer from severe over-segmentation; in such cases, the instance-independent D_{emb} metric provides a more robust signal.

Based on this comprehensive analysis, we calibrated general-purpose thresholds of $\tau_{FN} = 0.05$ and $\tau_{emb} = 0.5$. These fixed, pre-calibrated thresholds are used in all subsequent experiments to demonstrate a truly automated adaptation pipeline.

3.3. Adaptation Performance

Table 1 presents the main quantitative results of our method. On all three benchmarks, the baseline Cellpose models show a significant performance drop compared to supervised models. Our method, SelfAdapt, substantially closes this gap. When using the oracle "Test Max" as an upper bound, SelfAdapt improves the AP$_{0.5}$ by relative margins of 29.6% on LiveCell, 21.9% on TissueNet (Cytoplasm), and 26.0% on TissueNet (Nuclei).

Crucially, our automated label-free stopping criteria, using fixed thresholds calibrated on a separate representative run, capture a large portion of these potential gains. We validated this by applying the fixed thresholds ($\tau_{FN} = 0.05$, $\tau_{emb} = 0.5$) across all test runs. On the TissueNet (Nuclei) dataset, the same domain used for calibration, the FN and

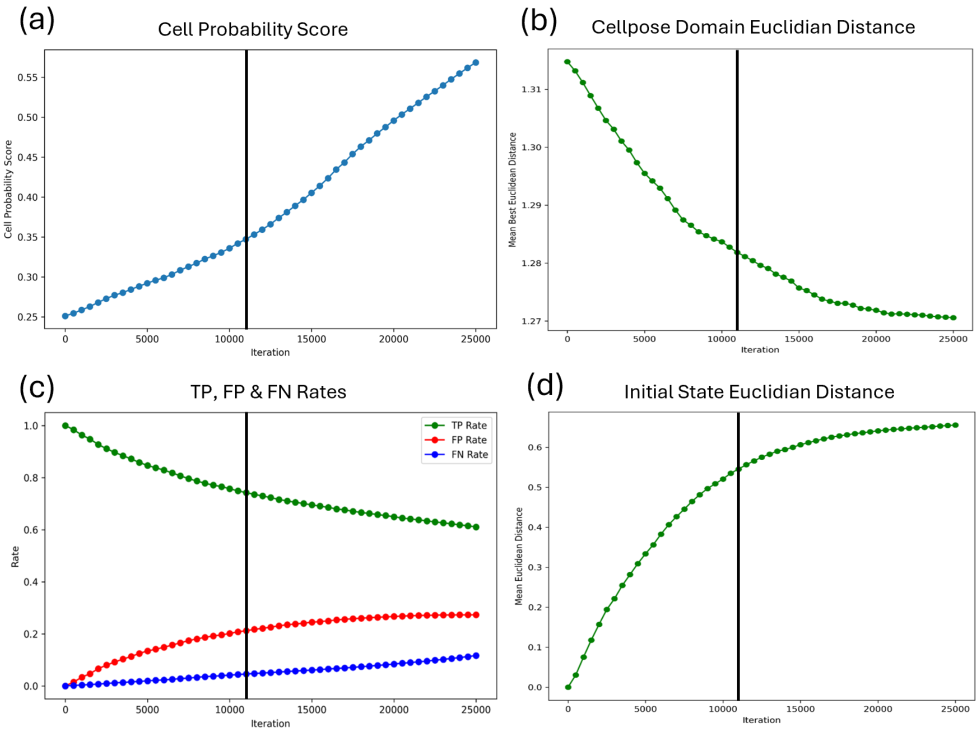

Figure 2. Analysis and selection of label-free early stopping criteria. We compare candidate metrics on a representative adaptation run (TissueNet Nuclei). The vertical black line indicates the iteration with the peak "oracle" performance on the test set. (a, b) Many intuitive metrics are unsuitable. Prediction confidence (a) increases monotonically as the model overfits, while proximity to the source domain (b) offers no clear stopping point and violates our source-free requirement. (c, d) In contrast, our chosen metrics provide a reliable signal. The drift in instance detection, measured by the FN Rate (c, blue line), and the drift in feature space, measured by the embedding distance D_{emb} (d), both produce monotonic curves with a clear, steady trend, making them ideal for thresholding.

Table 1. Performance comparison on LiveCell and TissueNet datasets. Our self-adaptation method improves both base and fine-tuned models, using three stopping criteria: False Negatives rate (FN), Embedding distance (Emb), and test-based maximum. Results show $\mathrm{AP}_{0.5}$ scores with standard deviations over three runs. For TissueNet (Nuclei), no fine-tuned model is available by [14]

Model	LiveCell (Cytoplasm)	TissueNet (Cytoplasm)	TissueNet (Nuclei)
Cellpose base model	.415	.324	.603
+ SelfAdapt(Early Stop: FN)	.480 ± .008	.394 ± .014	.759 ± .001
+ SelfAdapt(Early Stop: Emb)	.503 ± .003	.391 ± .013	.754 ± .003
+ SelfAdapt(Test Max)	.538 ± .002	.395 ± .015	.760 ± .001
Cellpose fine-tuned model	.695	.750	n/a
+ SelfAdapt(Early Stop: FN)	.693 ± .004	.752 ± .001	n/a
+ SelfAdapt(Early Stop: Emb)	.704 ± .001	.755 ± .000	n/a
+ SelfAdapt(Test Max)	.705 ± .002	.763 ± .000	n/a

Embedding Distance criteria achieve 99.4% and 96.2% of the maximum possible improvement, respectively.

More importantly, these fixed thresholds generalize remarkably well to entirely new datasets. For TissueNet (Cytoplasm), the FN and Embedding Distance criteria capture 98.6% and 94.4% of the maximum possible gain. For the visually distinct LiveCell dataset, the Embedding Distance criterion is particularly effective, achieving 71.5% of the maximum possible improvement, while the FN criterion still captures a substantial 52.8%. This demonstrates that our pre-calibrated thresholds can be used "off-the-shelf" to effectively adapt models to new domains without any labels

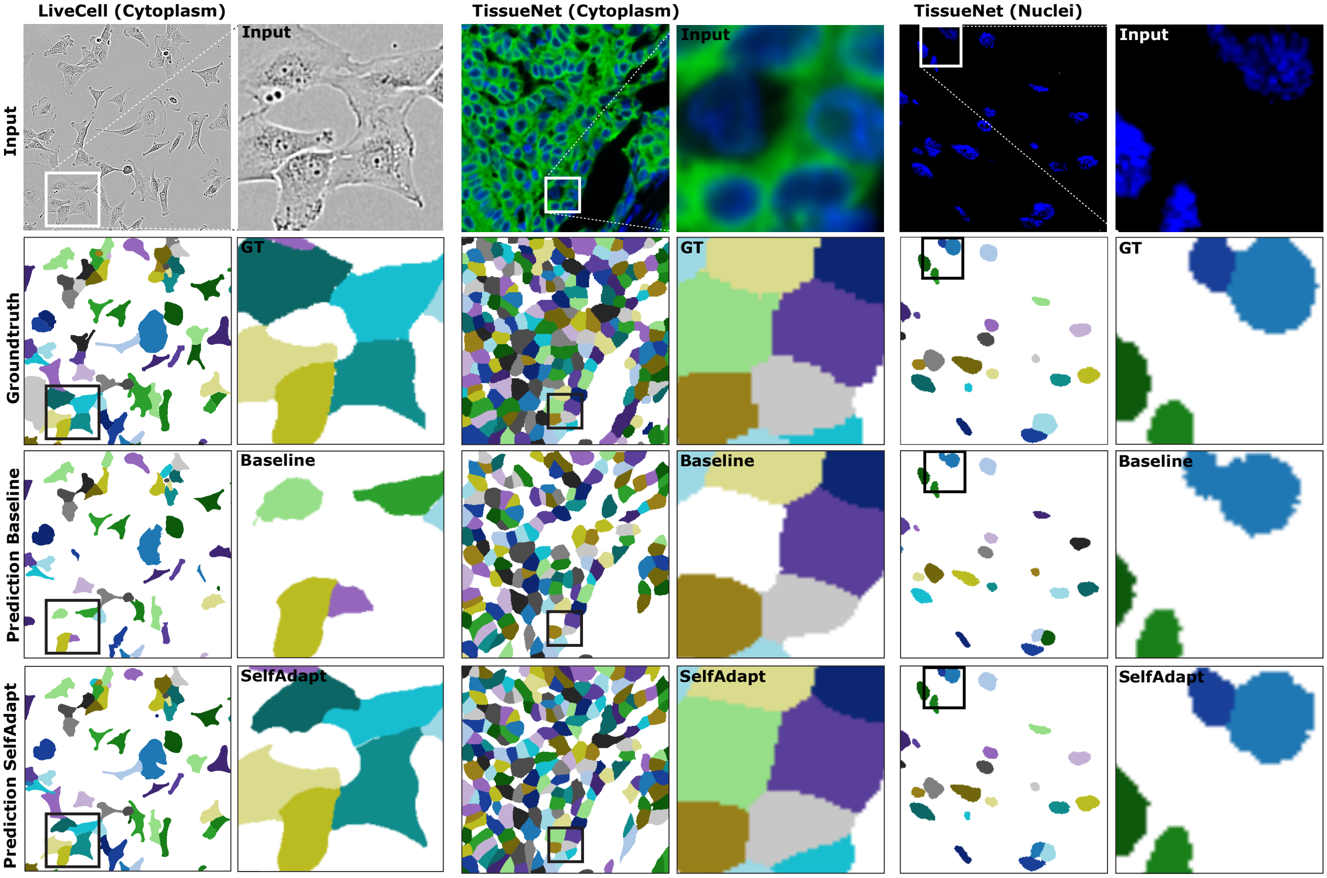

Figure 3. Qualitative comparison of segmentation results. Each column shows results on a different dataset: LiveCell (Cytoplasm), TissueNet (Cytoplasm), and TissueNet (Nuclei). For each dataset, we show the input image (top), ground truth segmentation (second row), baseline Cellpose predictions (third row), and predictions after self-adaptation (bottom row). Insets highlight cases where self-adaptation improves detection of previously missed cells (LiveCell, TissueNet Cytoplasm) and better separation of adjacent cells (TissueNet Nuclei).

Table 2. Ablation study examining the impact of key components in our method. Removing student augmentations or L2-SP regularization significantly impacts performance, while other components show more moderate effects.

Ablations				$AP_{0.5}$ split by dataset		
Confidence Filtering	Teacher TTA	L2-SP Loss	Student Augmentations	LiveCell (Cytoplasm)	TissueNet (Cytoplasm)	TissueNet (Nuclei)
✓	✓	✓	✓	.538 ± .002	.395 ± .012	.760 ± .001
✗	✓	✓	✓	.537 ± .004	.393 ± .006	.760 ± .001
✓	✗	✓	✓	.536 ± .001	.395 ± .013	.759 ± .001
✓	✓	✗	✓	.528 ± .005	.391 ± .021	.642 ± .009
✓	✓	✓	✗	.451 ± .005	.357 ± .004	.615 ± .002

or further tuning.

Interestingly, SelfAdapt can even further refine models that have already been fine-tuned with supervision (Table 1, bottom). For example, the Embedding Distance criterion further improves the fine-tuned LiveCell model, capturing 90.0% of the small remaining performance gap. While this demonstrates the robustness of our method, we note that in a supervised fine-tuning scenario, labeled validation data would typically be available, making automated label-free stopping less critical.

Figure 3 provides a qualitative view of these improvements. SelfAdapt enables the detection of cells missed by the baseline model and improves the separation of adjacent cells that were previously merged, corroborating the quantitative gains.

3.4. Ablation Study

To understand the contribution of each component of SelfAdapt, we conducted an ablation study (Table 2). The results highlight two critical components: strong student augmentations and L2-SP regularization. Removing student augmentations leads to a collapse in performance, as the student-teacher framework has no meaningful discrepancy to learn from. Removing L2-SP regularization is also highly detrimental, especially for the TissueNet (Nuclei) dataset, where it causes a 15.5% relative performance decrease. This confirms that constraining the model to stay close to its powerful initial weights is key for stable unsupervised adaptation. In contrast, confidence filtering and teacher test-time augmentation (TTA), while beneficial, have a more moderate impact.

4. Conclusion

We presented a practical method for unsupervised domain adaptation of cell segmentation models that enables immediate adaptation to new domains without requiring labels. Our approach combines augmentation consistency-based training with L2-SP regularization, demonstrating substantial improvements over baseline performance across multiple cellular image segmentation tasks. The success of L2-SP regularization suggests that for powerful generalist models like Cellpose, explicitly preventing the model from drifting too far from its robust initial weights is a key factor for stable adaptation, especially when learning from noisy pseudo-labels. The method proves effective not only for adapting base models but also for further improving supervised fine-tuned models.

SelfAdapt features a default, "off-the-shelf" early stopping criterion as determined via our experiments. At the same time, users can also leverage manual early stopping / snapshot selection and re-use respective inferred stopping criteria as they see fit.

We ship SelfAdapt as an easy-to-use extension of the Cellpose framework, requiring minimal setup while maintaining full compatibility with Cellpose's extensive functionality. This contribution aims to make domain adaptation more accessible to the biomedical research community, enabling improved cell segmentation across diverse imaging conditions without the need for additional annotations.

Acknowledgments

Funding: German Research Foundation (DFG) Research Training Group CompCancer (RTG2424); Synergy Unit of the Helmholtz Foundation Model Initiative.

References

[1] Anwai Archit and Constantin Pape. Probabilistic domain adaptation for biomedical image segmentation. *arXiv [cs.CV]*, 2023. 1, 2

[2] Christoffer Edlund, Timothy R Jackson, Nabeel Khalid, Nicola Bevan, Timothy Dale, Andreas Dengel, Sheraz Ahmed, Johan Trygg, and Rickard Sjögren. LIVECell-a large-scale dataset for label-free live cell segmentation. *Nat. Methods*, 18(9):1038–1045, 2021. 4

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 1

[4] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 597–613. Springer, 2016. 1

[5] Noah F Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Thomas Dougherty, Christine Camacho Fullaway, Brianna J McIntosh, Ke Xuan Leow, Morgan Sarah Schwartz, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology*, 40(4):555–565, 2022. 1, 2, 4

[6] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 1

[7] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018. 2

[8] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv [cs.NE]*, 2016. 2

[9] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning, ICML*, 3 (2):896, 2013. 2

[10] Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. *arXiv [cs.LG]*, 2018. 2

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv [cs.LG]*, 2017. 4

[12] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024. 4

[13] Ori Press, Ravid Shwartz-Ziv, Yann LeCun, and Matthias Bethge. The entropy enigma: Success and failure of entropy minimization. *arXiv [cs.CV]*, 2024. 3

[14] Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how

to train your own model. *Nat. Methods*, 19(12):1634–1641, 2022. 1, 2, 4, 5

[15] Joris Roels, Julian Hennies, Yvan Saeys, Wilfried Philips, and Anna Kreshuk. Domain adaptive segmentation in volume electron microscopy imaging. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1519–1522. IEEE, 2019. 1

[16] Josef Lorenz Rumberger, Jannik Franzen, Peter Hirsch, Jan-Philipp Albrecht, and Dagmar Kainmueller. ACTIS: Improving data efficiency by leveraging semi-supervised augmentation consistency training for instance segmentation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3792–3801. IEEE, 2023. 2

[17] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 265–273. Springer International Publishing, Cham, 2018. 1, 2

[18] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics (Basel)*, 13(11):1947, 2023. 1

[19] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv [cs.LG]*, 2020. 2

[20] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods*, 18(1):100–106, 2021. 1, 2, 4

[21] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv [cs.NE]*, 2017. 1, 2